Copyright © 2003 by The Journal of Bone and Joint Surgery, Incorporated

Current Concepts Review

User's Guide to the Surgical Literature: How to Use an Article About a Diagnostic Test

By Mohit Bhandari, MD, MSc, Victor M. Montori, MD, Marc F. Swiontkowski, MD, and Gordon H. Guyatt, MD, MSc

- The primary issues to consider in determining the validity of a diagnostic test study are how the authors assembled the patients and whether they used an appropriate reference standard for all patients to determine whether the patients did or did not have the target condition.
- Likelihood ratios are key to the interpretation of diagnostic tests as they link estimates of pretest probability to posttest probability.
- Sensitivity is the property of the test that describes the proportion of individuals with the disorder in whom the test result is positive.
- Specificity is the property of the test that describes the proportion of individuals without the disorder in whom the test result is negative.

Clinical Scenario

You are an orthopaedic surgeon who is asked to evaluate a sixty-five-year-old woman in the emergency department because of new-onset right hip pain that started one week ago. Seven months previously, the patient had had a right total hip arthroplasty for the treatment of osteoarthritis. The pain radiates to the thigh and buttocks. The patient reports that she slipped on a kitchen floor a few days ago but did not think that she had sustained a serious injury. In addition, she has been recovering from a sinus infection (a viral illness) for the past ten days. She is otherwise healthy except that she takes oral bisphosphonates for the treatment of osteoporosis.

On examination, she has a temperature of 39°C. She walks most comfortably with a flexed posture. The range of motion of the right hip is normal. There is no erythema or draining sinus over the right hip and thigh. Anteroposterior radiographs of the pelvis and the right hip reveal a press-fit acetabular component and a cemented femoral stem with no evidence of loosening. Laboratory evaluations show a white blood-cell count of 12.1 cells/μL, of which 85% are neutrophils. Blood cultures are negative.

You wonder whether the new onset of hip pain is the result of a soft-tissue injury, back pain radiating to the hip, prosthetic loosening that is not apparent on radiographs, or an infection of the hip joint. If the hip is truly infected, the patient will require an operative procedure for débridement of the wound and removal of the implants. While some of your colleagues would take all such patients to the operating room for exploration of the hip, you have been impressed by the number of cases in which you have found no infection. Because of such concerns, your practice is to routinely aspirate the hip in patients in whom an infection is suspected.

Just as you are thinking about placing your patient's name on the next day's procedures list for an aspiration, the result of the C-reactive protein test comes back as 8 mg/dL (normal, ≤10 mg/dL). This finding raises some question as to

This article is the fourth in a series designed to help the orthopaedic surgeon use the published literature in practice. In the first article in the series, we presented guidelines for making a decision about therapy and focused on randomized controlled trials. In the second article, we focused on evaluating nonrandomized studies that present information about a patient's prognosis. In the third article, we focused on systematic literature reviews. In this article, we address the use of articles about diagnostic tests in the care of surgical patients.

This reprint is made possible by an educational grant from Zimmer.

whether your patient actually has an infection. Unsure about the true utility of a C-reactive protein test in patients with a suspected infection, you decide to find a suitable article to clarify your concerns.

That evening, you conduct an Internet search to identify relevant articles to answer your question.

The Search

In preparation for your search, you formulate your question as follows: In patients with a previous total hip arthroplasty who are suspected of having an acute infection, what is the utility of a C-reactive protein test in diagnosing infection?

You have recently learned about the Clinical Queries function in PubMed, a quick way to narrow your search to identify articles that focus on diagnosis. Therefore, using the Clinical Queries search option in PubMed (http://www.ncbi.nlm.nih.gov/entrez/query/static/clinical.html), you choose a narrow scope search (specificity option) for articles on Diagnosis using the expression "C-reactive protein AND total joint arthroplasty." This search yields a single article entitled, "Prospective Analysis of Preoperative and Intraoperative Investigations for the Diagnosis of Infection at the Sites of Two Hundred and Two Revision Total Hip Arthroplasties," by Spangehl et al.[1]. A quick review of the abstract indicates that it will likely provide the information that you need. You obtain the article from your local hospital library.

Having decided on a relevant article, as is the case with other types of articles (therapy, prognosis, or harm), you should ask yourself three questions: (1) Are the results of the study valid? (2) What are the results? and (3) Will the results help me in caring for my patients? (Table I)[2].

Are the Results of the Study Valid?

Investigators studying a diagnostic test hope to establish the power of that test to differentiate between patients who have the target condition (i.e., the disease or health state—in this case, infection about a hip prosthesis) and those who are free of the target condition. Patients who are free of the target condition may be healthy or may have one of the competing diagnoses (in this case, hip trauma or pain radiating from the back). The credibility, believability, or validity of a study is only as good as the methods used in its conduct. The primary issues to consider in determining the validity of a diagnostic test study are how the authors assembled the patients and whether they used an appropriate reference standard for all patients to determine whether the patients did or did not have the target condition.

Was There Diagnostic Uncertainty?

How do you know whether the investigators chose a suitable population or whether their choice threatens the study's validity? The specific question to ask yourself is whether the surgeons who cared for the patients faced genuine diagnostic uncertainty. Tests are able to easily distinguish between severely affected and healthy patients (otherwise, they can easily be discarded from use). The reason for this excellent diagnostic performance relates to the minimal overlap between the test results for severely ill patients and the test results for healthy volunteers. However, clinicians are interested in using tests when there is diagnostic uncertainty, that is, when the test results for patients with the target condition are similar to the test results for patients without the target condition. In the latter group, diagnoses other than the target condition are responsible for the similarity of the test results between groups. Lijmer et al., in a report on bias in studies of diagnostic tests, demonstrated that studies involving patients with severe disease and healthy volunteers overestimated test performance threefold (relative diagnostic odds ratio = 3.0; 95% confidence interval, 2.0-4.5)[3].

TABLE I Guidelines for Evaluating Studies About a Diagnostic Test

Are the results of the study valid?
Primary guides
Did clinicians face diagnostic uncertainty?
Was there an independent, blind comparison with a reference standard?
Secondary guides
Did the results of the test being evaluated influence the decision to perform the reference standard?
Were the methods for performing the test described in sufficient detail to permit replication?
What are the results?
Are likelihood ratios of the test being evaluated or data necessary for their calculation provided?
Will the results help me in caring for my patients?
Will the reproducibility of the test result and its interpretation be satisfactory in my setting?
Are the results applicable to my patient?
Will the results change my management of the patient?
Will patients be better off as a result of the test?

For instance, the white blood-cell count will almost always be elevated in patients who present with an obvious hip infection that is associated with a draining sinus and pus in the joint. On the other hand, the white blood-cell count will almost never be elevated in healthy controls. However, its diagnostic utility is very poor in patients, like the one in the scenario described above, who may have early septic arthritis but who also may have another condition that elevates the white blood-cell count, such as viral pharyngitis, a urinary tract infection, or recent trauma.

The use of carcinoembryonic antigen for the detection of colorectal cancer provides a striking example of the variable utility of a diagnostic test in populations with different disease severity. Fletcher reported that carcinoembryonic antigen levels were elevated in thirty-five of thirty-six patients with established cancer and were much lower in patients without cancer[4]. However, in an another study in which carcinoembryonic antigen testing was applied to patients with less-advanced stages of

This reprint is made possible by an educational grant from Zimmer.

colorectal cancer, the test results were similar enough to those in patients without cancer that the ability of the test to distinguish the two groups declined[5]. Accordingly, the use of carcinoembryonic antigen in the diagnosis of cancer was abandoned.

Spangehl et al. included a wide spectrum of patients with low, moderate, and high levels of clinical suspicion of infection[1]. We can therefore conclude that the authors assembled an appropriate spectrum of patients.

It is important to recognize that the predictive value of a test will change with changes in the prevalence of the disease spectrum already discussed. Consider the following situation. When a test to diagnose influenza infection (the common flu virus) is used during an influenza season, positive test results are more likely to truly indicate cases of influenza than they are when the same test is used in the same community during the off-season. This difference occurs because there are more cases (i.e., a higher prevalence) of influenza during the influenza season and not because the diagnostic properties of the test have changed.

Was There an Independent Comparison with a Reference Standard?

The accuracy of a diagnostic test is best determined by comparing it with the truth. Truth about whether the disease is present is usually defined by the presence or absence of a pathological finding that represents the condition (i.e., an essential lesion). A reference standard that uses that pathological finding is most desirable. Conversely, a reference standard that does not use an essential lesion is at risk of miscategorizing patients. Therefore, judgment should be used to decide whether the chosen reference is appropriate.

Accordingly, readers must make sure that the investigators have applied independently both the test under investigation and an appropriate *reference standard* (such as biopsy, surgery, autopsy, or long-term follow-up) to every patient. By *independent*, we mean that the individual interpreting the reference standard should be unaware of (or blind to) the results of the test and that the individual interpreting the test should be unaware of the results of the reference standard. To the extent that this blinding is not achieved, the investigation is likely to overestimate the diagnostic power of the test. In the study by Lijmer et al.[3], lack of blinding resulted in a significant overestimation of the test performance (relative diagnostic odds ratio = 1.3; 95% confidence interval, 1.0-1.9) ($p < 0.05$).

For example, surgeons who detect a hip fracture with use of nuclear bone-scanning or magnetic resonance imaging are more likely to identify a previously undetected fracture line on plain radiographs. In one study evaluating the use of plain radiography and magnetic resonance imaging for the detection of osteonecrosis following a hip fracture, the investigators did not report independent assessments of plain radiographs and magnetic resonance images[6]. Thus, the investigators who identified changes on magnetic resonance images at two months may have been more suspicious of the findings on the plain radiographs, which initially appeared normal but ultimately were classified as abnormal.

Another way in which a lack of independence can be misleading is if the test under evaluation is a component of the reference standard. For example, in one study investigating the utility of the serum and urinary amylase test in the diagnosis of pancreatitis, the investigators constructed a reference standard that consisted of a series of tests, including the serum and urinary amylase test[7]. This incorporation of the test under evaluation into the reference standard is likely to overestimate the utility of the test. Thus, clinicians should make sure that the test under evaluation and the reference standard are independent of each other.

In the study by Spangehl et al., all patients underwent measurement of the C-reactive protein level and testing to determine the presence or absence of infection. The authors did not describe clearly whether the assessments were performed in an independent and blinded fashion[1]. The investigators defined infection as the presence of an open or draining sinus communicating with the hip joint, the detection of purulent fluid within the joint during surgical exploration, or a positive result on at least three other investigations (intraoperative culture, preoperative aspiration, frozen-section analysis, determination of the C-reactive protein level, and determination of the erythrocyte sedimentation rate). The inclusion of the diagnostic test in question (the C-reactive protein test) as a component of this reference standard raises a serious concern. This *incorporation bias* may spuriously increase the apparent utility of the test.

Having asked the most critical questions that assist in the determination of study validity, you can further reduce your chances of being misled by asking an additional question.

Did the Results of the Test Being Evaluated Influence the Decision to Perform the Reference Standard?

The properties of a diagnostic test will be distorted if the results of the test influence the decision to carry out the reference standard. This situation, called *verification bias*[8,9] or *workup bias*[10,11], applies when, for example, investigators only conduct further evaluation with the reference standard for patients who have a positive test result and assume that those who have a negative test result do not have the target condition. In practice, this leads to an overly sanguine estimation of the ability of the test being evaluated to differentiate between patients who have the target condition and those who do not. In the study by Lijmer et al., the test performance was overestimated twofold in studies in which different reference standards were used for patients who had the target condition and those who did not (relative diagnostic odds ratio = 2.2; 95% confidence interval, 1.5-3.3)[3].

Generally, if a test is invasive, surgeons will be less likely to apply the reference standard (i.e., surgical biopsy) when the probability of disease is low. Verification bias occurred in a study of the diagnostic utility of fine-needle aspiration biopsy in the determination of malignancy in patients with nodular thyroid disease[12]. Patients who had benign lesions on fine-needle aspiration biopsy did not have surgical resection of the thyroid nodule for definitive pathological diagnosis, whereas those

This reprint is made possible by an educational grant from Zimmer.

TABLE II Likelihood Ratios for a Positive and Negative C-Reactive Protein Test*

C-Reactive Protein Test	Periprosthetic Infection		Total
	Yes	No	
Positive (>10 mg/L)	25 True Positive (a)	9 False Positive (b)	34
Negative (≤10 mg/L)	1 False Negative (c)	107 True Negative (d)	108
Total	26	116	

Likelihood ratio (for positive test): (a/[a + c])/(b/[b + d]) = sensitivity/(1 – specificity) = (25/26)/(9/116) = 0.96/0.077 = 12.5.
Likelihood ratio (for negative test): (c/[a + c])/(d/[b + d]) = (1 – sensitivity/specificity) = (1/26)/(107/116) = 0.038/0.92 = 0.041.
Sensitivity: a/(a + c) = 25/26 = 96%.
Specificity: d/(b + d) = 107/116 = 92%.
Positive predictive value: a/(a + b) = 25/34 = 74%.
Negative predictive value: d/(c + d) = 107/108 = 99%.
Accuracy: (a + d)/(a + b + c + d) = 132/142 = 93%.
Prevalence: (a + c)/(a + b + c + d) = 26/142 = 18%.

*The data are from the study by Spangehl et al.[1].

who had malignant or uncertain lesions on fine-needle aspiration biopsy underwent a further reference standard examination with surgical resection and pathological analysis. That study is likely to have overestimated the power of the test in excluding malignancy.

Verification bias was also a potential problem in the landmark study of the value of the ventilation-perfusion lung scan in the diagnosis of pulmonary embolism (the PIOPED study)[13]. Patients whose ventilation-perfusion scans were interpreted as "normal/near normal" and "low probability" were less likely to undergo pulmonary angiography than those with more positive ventilation-perfusion scans; specifically, 69% of the patients in the former group and 92% of those in the latter group underwent angiography[13]. This finding is not surprising as clinicians might be reluctant to subject patients who have a low probability of pulmonary embolism to the risks of angiography. In this case, however, the investigators dealt successfully with the bias by constructing an alternative reference standard for patients who did not undergo angiography. They followed these untreated patients for one year to ensure that they remained free of evidence of pulmonary embolism during this period of time.

The methods section of the article by Spangehl et al.[1] indicates that all patients underwent frozen-section analysis as well as intraoperative gram-staining and culture of specimens from the surgical site. Thus, the results of the C-reactive protein test did not influence the decision to conduct reference standard investigations in these patients. What is less clear is whether the investigators interpreting the reference standard had access to the results of the C-reactive protein test.

What Are the Results?

The starting point for any diagnostic process is to determine the probability that the target disease is present in a given patient group before the next diagnostic test is performed. Let us consider two patients: (1) a sixty-five-year-old woman with diabetes who presents six months after total hip arthroplasty with a fever, an elevated white blood-cell count, and a painful hip with an erythematous wound, and (2) a sixty-year-old otherwise healthy woman who presents one year after arthroplasty with intermittent hip pain, normal findings on physical examination, and an elevated white blood-cell count. Most surgeons would consider the probability of an infection about the prosthesis to be different for these two patients. The probability, referred to as the pretest probability, of periprosthetic infection in the sixty-five-year-old patient with hip pain and fever is much higher than the probability of infection in the sixty-year-old patient even before additional diagnostic tests are conducted.

How can surgeons estimate pretest probability? Literature on the probability of disease given a certain presentation (for example, reports discussing the probability of infection in patients presenting with pain and fever after arthroplasty), similar data derived from the hospital's registry, and a surgeon's clinical experience and intuition can help that surgeon to estimate pretest probability. Other information that can be used to estimate pretest probability can be found in studies evaluating the utility of a diagnostic test. For instance, in the study by Spangehl et al.[1], 17% (thirty-five) of the 202 hips were found to be infected.

Returning to your patient, you can use the history and clinical examination to arrive at a pretest probability (that is, the probability of infection before the result of the C-reactive protein test was obtained). Your patient's elevated white blood-cell count and fever are consistent with her recent viral infection. However, the new-onset hip pain raises concern that she may have a periprosthetic infection. The wound is neither erythematous nor warm to the touch. Indeed, this patient is similar to an average patient in the study by Spangehl et al.[1]. On the basis of this information, you estimate that your patient has a 20% probability of a periprosthetic infection.

The next step is to decide how the results of the C-reactive protein test change your estimate of the probability of infection. In other words, surgeons should be interested in the char-

This reprint is made possible by an educational grant from Zimmer.

acteristic of the test that indicates the direction and magnitude of this change. This characteristic of the test is termed the *likelihood ratio*[2]. The likelihood ratio (LR) is the characteristic of the test that links the pretest probability to the posttest probability (that is, the probability of the target condition after the test results are obtained).

What Are the Likelihood Ratios Associated with the Test Results?

Table II presents results from the study by Spangehl et al.[1] (although not in the way that the authors presented them). There were twenty-five patients who had a proven infection and 107 patients in whom infection was ruled out. For all patients, the C-reactive protein level was classified as positive (>10 mg/L) or negative (≤10 mg/L). How likely is a negative C-reactive protein test among patients who have a periprosthetic infection? Table II reveals that the C-reactive protein level was normal in one (4%) of twenty-six patients with an infection and in 107 (92%) of 116 patients without an infection. The ratio of these two proportions (0.04/0.92) is the likelihood ratio for a negative C-reactive protein test and is equal to 0.043. Thus, a negative C-reactive protein test is twenty-three times (that is, 1/0.043 times) less likely to occur in patients with a periprosthetic infection than in those without an infection. Alternatively, a positive C-reactive protein test is 12.5 times more likely to occur in patients with a periprosthetic infection than in those without an infection (Table II).

How can we use the likelihood ratio? The likelihood ratio tells us how much the pretest probability increases or decreases. For instance, a likelihood ratio of 1.0 will not change the pretest probability, whereas a likelihood ratio of >1 will increase it. A rough guide to the interpretation of likelihood ratios is as follows: likelihood ratios of >10 or <0.1 generate large and often conclusive changes in the posttest probability, likelihood ratios from >5 to 10 or from 0.1 to 0.2 generate moderate shifts in posttest probability, likelihood ratios from >2 to 5 or from >0.2 to 0.5 generate small (but sometimes important) changes in probability, and likelihood ratios from >1 to 2 or from >0.5 to 1 alter posttest probability to a small degree[2].

Having determined the likelihood ratios, how do we use them to link the pretest probability to the posttest probability? A simple but tedious calculation converts the pretest probability to pretest odds (odds = probability/[1 − probability]). The clinician can then multiply the pretest odds by the likelihood ratio to obtain the posttest odds. With use of another calculation, the posttest odds can be converted back to posttest probability (probability = odds/[1 + odds]).

To save time and avoid computations, Fagan proposed a nomogram for converting pretest probability to posttest probability with use of likelihood ratios[14]. The clinician obtains the posttest probability by placing a straight edge that aligns the pretest probability to the likelihood ratio for the diagnostic test. For your patient who has a pretest probability of 20% on the basis of history and clinical examination and a negative C-reactive protein test (LR = 0.04), the posttest probability is 1%. If the C-reactive protein test had been positive (LR = 12.5), then the posttest probability of an periprosthetic infection would have increased to 76%.

Table III illustrates how this approach would be applied to the two patients presented earlier: the sixty-five-year-old woman with hip pain and overt signs of infection (pretest probability = 80%) and the sixty-year-old woman with hip pain but no fever (pretest probability = 10%). Formally, new knowledge (posttest probability) that is derived from the revision of previous knowledge (pretest probability) on the basis of new information (likelihood ratio) is an application of Bayes theorem to diagnosis.

As is evident from the above examples, the use of likelihood ratios is key to the interpretation of diagnostic tests. However, many studies present the properties of diagnostic tests in less clinically useful terms: sensitivity and specificity.

Sensitivity, Specificity, and Predictive Value (see Table II)

Sensitivity is the property of the test that describes the proportion of patients with the disorder in whom the test result is positive. Specificity is the property of the test that describes the proportion of patients without the disorder in whom the test result is negative. Using the rules provided in Table II, we can calculate the sensitivity and specificity of the C-reactive protein test in detecting infection. To calculate sensitivity, we divide the total number of patients who had a proven infection and a positive test (true positives; n = 25) by the total number of patients who had a proven infection (true positives + false negatives; n = 26). Thus, the sensitivity is 96%. To calculate specificity, we divide the total number of patients who had a negative C-reactive protein test (true negatives; n = 107) by the total number of patients who had no infection (true negatives + false positives; n = 116). Therefore, the specificity is 92%.

Tests with high sensitivity are useful for ruling out disease, and tests with high specificity are useful for ruling in disease. For example, since almost all patients with a scaphoid

TABLE III Pretest Probabilities, Likelihood Ratios, and Posttest Probabilities

Pretest Probability *(%)*	Likelihood Ratio*	Posttest Probability *(%)*
Negative test		
80 (high probability)	0.04	14
50	0.04	3.8
30	0.04	1.8
10 (low probability)	0.04	0.4
Positive test		
80 (high probability)	12.5	98
50	12.5	93
30	12.5	84
10 (intermediate probability)	12.5	60

*As determined on the basis of the result of the C-reactive protein test.

This reprint is made possible by an educational grant from Zimmer.

fracture suffer from anatomical snuffbox tenderness (a highly sensitive test), the absence of such tenderness virtually rules out a scaphoid fracture[15]. In patients with a neck injury, the absence of five clinical features (midline cervical tenderness, focal neurological deficit, impaired alertness, intoxication, and history of a distraction injury) reduces the probability of an important cervical spine injury to <1%[16]. In patients suspected of having a full-thickness rotator cuff tear, a normal ultrasound rules out a full-thickness tear because ultrasonography has a sensitivity of 100%[17].

The three examples cited above are all situations in which a high-sensitivity test, if negative, can rule out a target condition. The posterior drawer test for the diagnosis of posterior cruciate ligament injury is highly specific. Rubinstein et al. conducted a study to determine the diagnostic utility of the posterior drawer test among a varied population of patients, including those with normal knees, those with anterior-cruciate-deficient knees, and those with posterior-cruciate-deficient knees[18]. Among blinded assessors, a specificity of 99% was reported. Thus, a positive result on the posterior drawer test makes the diagnosis of posterior cruciate ligament injury virtually certain.

Sensitivity and specificity have drawbacks. In calculating sensitivity and specificity, important information is often discarded to collapse the data to fit the 2 × 2 table format. Moreover, multiple recalculations of sensitivity and specificity are often necessary at each potential cut point (or division) when one is considering a continuous variable (for example, blood pressure) or a test result that is reported as one of a number of categories (such as a high, intermediate, or low-probability ventilation-perfusion scan). Finally, there is no convenient nomogram that allows us, with knowledge of sensitivity, specificity, and a particular test result, to convert pretest probability to posttest probability. However, one can translate these measures into likelihood ratios. Similar drawbacks affect the calculation of predictive values (Table II).

Will the Results Help Me in Caring for My Patients?

Having assessed the validity of the article and performed the necessary simple calculations to understand its results, you can ask yourself whether these results will help you in caring for your patient.

The value of a diagnostic test often depends on its reproducibility when applied to patients. If a test requires much interpretation (e.g., electrocardiograms or pathological specimens) or involves the use of laboratory assays (e.g., stains or biochemical assays), variation in test results can occur. If a study indicates that a test is highly reproducible, two possibilities are likely: either the test is quite simple and easy to apply to patients or the investigators involved in the study were highly skilled in applying the diagnostic test to the study patients. If the latter is true, the diagnostic test may not be useful in a setting in which nonskilled interpretation of the test is likely to occur.

Another important issue to consider is the similarity of your patient to those in the study. The properties of a diagnostic test can change with different disease severities (see the discussion on the use of an appropriate spectrum, above). For instance, the test may not perform as well in a community practice, where less complicated cases will have to be distinguished from multiple competing diagnoses. On the other hand, in the study by Spangehl et al.[1], the patients were assessed in a referral practice setting (a university hospital). In that setting, surgeons were more likely to encounter patients with more severe or complicated disease in whom the diagnostic test (the C-reactive protein level) was likely to perform better (likelihood ratio >>1). In that setting, alternative diagnoses may have already been explored and ruled out. Likelihood ratios tend to move away from the value of 1 when all patients who have the target disorder have severe disease, and they tend to move toward the value of 1 when all patients who have the target disorder have mild disease[2]. In general, however, if you practice in a similar setting to that presented in the study and your patient meets the study eligibility criteria, you can be confident in applying the results of the study to your patient.

Once you have decided that the results are, in fact, applicable to your patient, you must decide whether they will change your management of the patient. Before making any decisions, you must have a sense of what probabilities would confirm or refute the target diagnosis. For example, suppose you are willing to proceed with débridement and implant removal without further testing in patients who have a ≥85% probability of infection (realizing that you will be operating on 15% of patients unnecessarily). Moreover, suppose you are willing to reject the diagnosis of infection if the test probability is ≤10%. In the sixty-five-year-old woman with hip pain and overt signs of infection (pretest probability, 80%) and a negative C-reactive protein test, the posttest probability of periprosthetic infection would be 14% and you would proceed with further testing (e.g., hip aspiration) before abandoning infection as a diagnosis. However, in the sixty-year-old afebrile woman with hip pain (pretest probability, 10%) and a negative C-reactive protein test, the posttest probability of infection would be nearly 0% and you would not conduct further testing for periprosthetic infection. You may wish to apply different numbers here; the treatment and test thresholds are a matter of values (ideally, the patient's values) and they differ among conditions depending on the risks of therapy (i.e., if the therapy is associated with severe side effects, you may want to be more certain of your diagnosis before recommending it) and the danger of the disease if left untreated (i.e., if the danger of missing the disease is high—as it is in the case of pulmonary embolism—you may want your posttest probability to be very low before abandoning diagnostic testing) (Fig. 1).

Finally, you can ask yourself if your patient will be better off having had the test. A test becomes more valuable when it has acceptable risks, the target disorder has major consequences if left untreated, and the target disorder can be readily treated if diagnosed. C-reactive protein testing poses minimal

This reprint is made possible by an educational grant from Zimmer.

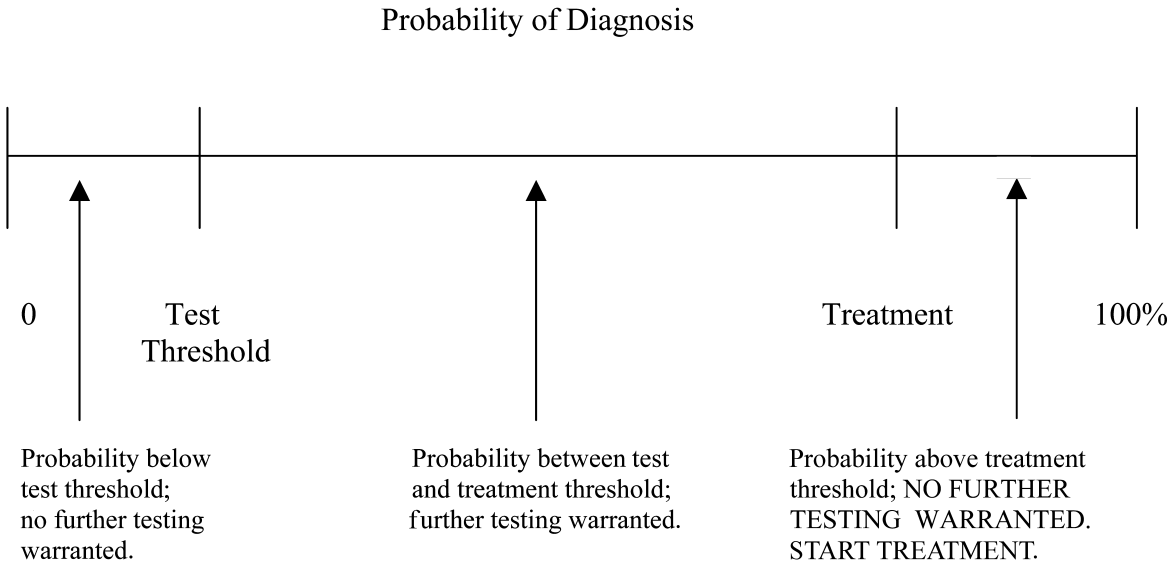

Fig. 1

Diagram illustrating diagnostic thresholds.

risk to the patient and may be extremely valuable for ruling in or ruling out infection—a complication of total hip arthroplasty that is devastating if left untreated.

Resolution of the Scenario

The patient in the scenario at the beginning of this report had a pretest probability of infection of 20%. Her negative C-reactive protein test (likelihood ratio, 0.04) decreased her probability of infection to 1%. The patient did not undergo a surgical procedure but required close follow-up. At the two-week follow-up appointment, the white blood-cell count was normal and the patient was afebrile. Further examination of radiographs and computed tomographic scans of the lumbar spine revealed right lateral recess stenosis.

Conclusion

Application of the guides presented in this article can allow surgeons to critically assess studies about a diagnostic test. Surgeons are continuously exposed to a variety of new and innovative diagnostic tests and to the studies describing their diagnostic properties. Determining the validity of these studies, the study results, and the applicability of these results to your patients are three fundamental steps toward choosing and interpreting diagnostic tests.

Mohit Bhandari, MD, MSc
Victor M. Montori, MD
Gordon H. Guyatt, MD, MSc
Department of Clinical Epidemiology and Biostatistics, McMaster University Health Sciences Center, 1200 Main Street West, Hamilton, ON L8N 3Z5, Canada. E-mail address for M. Bhandari: bhandari@sympatico.ca

Marc F. Swiontkowski, MD
Department of Orthopaedic Surgery, University of Minnesota, Box 492, Delaware Street N.E., Minneapolis, MN 55455

The authors did not receive grants or outside funding in support of their research or preparation of this manuscript. They did not receive payments or other benefits or a commitment or agreement to provide such benefits from a commercial entity. No commercial entity paid or directed, or agreed to pay or direct, any benefits to any research fund, foundation, educational institution, or other charitable or nonprofit organization with which the authors are affiliated or associated.

References

1. **Spangehl MJ, Masri BA, O'Connell JX, Duncan CP.** Prospective analysis of preoperative and intraoperative investigations for the diagnosis of infection at the sites of two hundred and two revision total hip arthroplasties. *J Bone Joint Surg Am.* 1999;81:672-83.
2. **Jaeschke R, Guyatt G, Sackett DL.** Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. *JAMA.* 1994;271:389-91.
3. **Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM.** Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA.* 1999;282:1061-6.
4. **Fletcher RH.** Carcinoembryonic antigen. *Ann Intern Med.* 1986;104:66-73.
5. **Thomson DM, Krupey J, Freedman SO, Gold P.** The radioimmunoassay of circulating carcinoembryonic antigen of the human digestive system. *Proc Natl Acad Sci USA.* 1969;64:161-7.
6. **Kawasaki M, Hasegawa Y, Sakano S, Sugiyama H, Tajima T, Iwasada S, Iwata H.** Prediction of osteonecrosis by magnetic resonance imaging after femoral neck fractures. *Clin Orthop.* 2001;385:157-64.
7. **Kemppainen EA, Hedstrom JI, Puolokkainen PA, Sainio VS, Haapiainen RK, Perhoniemi V, Osman S, Kivilaakso EO, Stenman UH.** Rapid measurement

This reprint is made possible by an educational grant from Zimmer.

of urinary trypsinogen-2 as a screening test for acute pancreatitis. *N Engl J Med.* 1997;336:1788-93.

8. **Begg CB, Greenes RA.** Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics.* 1983;39:207-15.
9. **Gray R, Begg CB, Greenes RA.** Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Med Decis Making.* 1984;4:151-64.
10. **Ransohoff DF, Feinstein AR.** Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med.* 1978;299:926-30.
11. **Choi BC.** Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. *J Clin Epidemiol.* 1992;45:581-6.
12. **Hamming JF, Goslings BM, van Steenis GJ, van Ravenswaay Claasen H, Hermans J, van de Velde CJ.** The value of fine-needle aspiration biopsy in patients with nodular thyroid disease divided into groups of suspicion of malignant neoplasms on clinical grounds. *Arch Intern Med.* 1990;150:113-6.
13. **The PIOPED Investigators.** Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). *JAMA.* 1990;263:2753-9.
14. **Fagan TJ.** Letter: nomogram for Bayes theorem. *N Engl J Med.* 1975;293:257.
15. **Parvizi J, Wayman J, Kelly P, Moran CG.** Combining the clinical signs improves diagnosis of scaphoid fractures. A prospective study with follow-up. *J Hand Surg [Br].* 1998;23:324-7.
16. **Hoffman JR, Mower WR, Wolfson AB, Todd KH, Zucker MI.** Validity of a set of clinical criteria to rule out injury to the cervical spine in patients with blunt trauma. National Emergency X-Radiography Utilization Study Group. *N Engl J Med.* 2000;343:94-9.
17. **Teefey SA, Hasan SA, Middleton WD, Patel M, Wright RW, Yamaguchi K.** Ultrasonography of the rotator cuff. A comparison of ultrasonographic and arthroscopic findings in one hundred consecutive cases. *J Bone Joint Surg Am.* 2000;82:498-504.
18. **Rubinstein RA Jr, Shelbourne KD, McCarroll JR, VanMeter CD, Rettig AC.** The accuracy of the clinical examination in the setting of posterior cruciate ligament injuries. *Am J Sports Med.* 1994;22:550-7.

This reprint is made possible by an educational grant from Zimmer.